SLiMSearch 2.0: biological context for short linear motifs in proteins

Norman E. Davey¹, Niall J. Haslam², Denis C. Shields²* and Richard J. Edwards³

¹Structural and Computational Biology Unit, European Molecular Biology Laboratory, 69117 Heidelberg, Germany, ²UCD Complex and Adaptive Systems Laboratory, UCD Conway Institute, and School of Medicine and Medical Sciences, University College Dublin, Dublin 4, Ireland and ³School of Biological Sciences, University of Southampton, Southampton, UK

Received February 12, 2011; Revised April 26, 2011; Accepted May 5, 2011

ABSTRACT

Short, linear motifs (SLiMs) play a critical role in many biological processes. The SLiMSearch 2.0 (Short, Linear Motif Search) web server allows researchers to identify occurrences of a user-defined SLiM in a proteome, using conservation and protein disorder context statistics to rank occurrences. User-friendly output and visualizations of motif context allow the user to quickly gain insight into the validity of a putatively functional motif occurrence. For each motif occurrence, overlapping UniProt features and annotated SLiMs are displayed. Visualization also includes annotated multiple sequence alignments surrounding each occurrence, showing conservation and protein disorder statistics in addition to known and predicted SLiMs, protein domains and known post-translational modifications. In addition, enrichment of Gene Ontology terms and protein interaction partners are provided as indicators of possible motif function. All web server results are available for download. Users can search motifs against the human proteome or a subset thereof defined by Uniprot accession numbers or GO term. The SLiMSearch server is available at: http://bioware.ucd.ie/slimsearch2.html.

INTRODUCTION

The purpose of the SLiMSearch (Short, Linear Motif Search) web server is to allow researchers to identify novel occurrences of user-defined Short Linear Motifs (SLiMs) in a set of sequences. SLiMs, also referred to as linear motifs or minimotifs, are functional microdomains that play a central role in many diverse biological pathways (1) through post-translational modification (including cleavage), subcellular localization and ligand binding (2). Once a SLiM has been defined, finding matches in a given set of protein sequences is a fairly trivial task. Several web-based methods to discover novel instances of known SLiMs are available, including ELM (2), MnM (3), SIRW (4) ScanProsite (5) and QuasiMotifFinder (6), which generally utilize databases of known motif patterns to search query protein sequences supplied by the user.

While finding matches is trivial, however, interpreting their biological significance is far from easy. Stochastic occurrences of small, degenerate motifs are common; distinguishing real occurrences from the background of random motif hits remains the greatest challenge in a priori motif discovery. One approach is to simply filter out motifs that are likely to occur numerous times by chance—ScanProsite (5), for example, has an option to ‘Exclude motifs with a high probability of occurrence’, while QuasiMotifFinder (6) uses the background occurrence of motifs in PfamA families (7) to assess the significance of hits. These strategies work well for longer, family descriptor motifs [such as are found in the Prosite database (8) used by both ScanProsite and QuasiMotifFinder] but are not so useful for SLiMs because of their tendency to occur by chance. Instead, additional contextual information such as sequence conservation (3,6,9,10), structural context (3,11) or even biological keywords (4) can be used to assess the likelihood of true functional significance for putatively functional sites.

Most motif search tools rely on pre-existing motif libraries, such as ELM (2), MnM (3) or Prosite (8). Those that permit users to define their own motifs, such as ScanProsite (5), are generally lacking the contextual information required to aid functional inference. Recent developments in de novo motif discovery has given rise to a number of tools that are capable of predicting entirely novel SLiMs from sets of protein sequences [e.g. PRATT (12), MEME (13), Dilmot (14), SLiMDisc (15), ...]
SLiMFinder (16) and FIRE-pro (17)]. Although SLiMFinder (16) estimates the statistical significance of returned motif predictions, correcting for biases introduced by evolutionary relationships within the data, assessing the ‘biological’ significance of predicted SLiMs remains challenging. One approach is to compare candidate SLiMs to existing motif libraries to identify similarities to previously known motifs (18). When a genuinely novel motif is predicted, however, knowledge of existing motifs is of limited use. Instead, it is useful to be able to establish the background distribution of occurrences of the novel motif, utilizing contextual information to help screen out the inevitable spurious chance matches.

We recently made our powerful de novo SLiM discovery tool, SLiMFinder (16), available as a web server (19). To aid interpretation of SLiMFinder results, we made a new tool available, SLiMSearch, which allows users to search protein data sets with user-defined motifs, including motif prediction output from SLiMFinder (20). SLiMSearch utilized the same sequence context assessment as SLiMFinder, enabling results to be masked or ranked based on the important biological indicators of sequence conservation and structural disorder (10,21) and features the same SLiMChance algorithm for assessing statistical overrepresentation of SLiM occurrences, correcting for biases introduced by evolutionary relationships within the data (16). Like SLiMFinder, SLiMSearch was optimized for small protein data sets. In this article, we describe a complementary server, SLiMSearch 2.0, which is optimized for searches of a whole proteome.

SLiMSearch 2.0 replaces SLiMChance data set probabilities with individual likelihoods for each motif instance that permit the ranking of many motif occurrences and helps separate putative functional instances from the background of stochastic occurrences. A comprehensive study of the Eukaryotic Linear Motif (ELM) database by Fuxreiter et al. (22) found that SLiMs are more likely to be found in disordered regions, while Chica et al. (9) found that conserved motifs are more likely to be true positives. Our previous work with both discovery of new instances of known motifs and of novel motifs shows that, motifs in disordered regions and conserved motifs are typically (but not always) more likely to be true positives (10). Therefore, we encourage the use of an optional disorder filter and we present the results ranked according to conservation. Enrichment scores for motif counts are calculated (i) versus reversed/shuffled variants of the motif, (ii) for Gene Ontology (GO) terms (23) and (iii) for known BioGRID interactors of individual hub proteins (24). In addition to identifying individual occurrences of known motifs, therefore, SLiMSearch 2.0 can indicate possible functional significance for entirely novel motifs. Input, output and results visualizations are fully compatible with our existing SLiM analysis web servers, SLiMDisc (25), CompariMotif (18), SLiMFinder (19) and SLiMSearch 1.0 (20), providing a suite of integrated tools for analysing these biologically important sequence features.

**THE SLI MS EARCH 2.0 ALGORITHM**

SLiMSearch 2.0 performs a motif regular expression search against a proteome allowing restriction of considered sequences to set of proteins or a given GO term. Features include annotation of overlapping sequence annotation and calculation of global and local motif statistics and attributes.

**Pre-formatted database**

To speed up motif attribute calculations, pre-computed databases for each proteome are used. The current release has only Human UniProt release v1.37 (Aug 2010) (26); however, more model proteomes will be added as data is computed. Two pre-computed conservation scores are calculated for each protein in the proteome, a column-based tree-weighted conservation score (WCS) (9) and a relative local conservation (RLC) metric (10). Homologues for each sequence are identified using a BLAST search against a database of 70 complete Ensembl proteomes (Ensembl 59, October 2010, 69 Metazoan proteomes and Saccharomyces cerevisiae) (27) and orthologues are predicted using GOPHER (default options) (25). Predicted orthologues are aligned by MAFFT (28) and used to calculate conservation scoring metrics on a residue-by-residue basis. Disorder scores for each residue are calculated using IUPred (default options (21)).

Several features of interest are also preformatted for rapid querying: (i) Domain data from Pfam (29); (ii) structure data from PDB (30); (iii) experimentally validated motifs from the ELM database (2); and (iv) SNP and modification data from UniProt annotation (26).

**Scoring**

The IUPred disorder score, IUP, of the motif is calculated as the mean disorder score across the defined (non-wildcard) residues. The WCS of a motif is calculated similarly. SLiMSearch 2.0 extends the RLC score to return a probability. Based on the assumption, consistent with empirical observation, that the RLC scores for a residue are normally distributed (10), the RLC of a residue is converted into a probability, \( P_{\text{RLC}} \), using the Gaussian Cumulative Distribution Function (CDF). The relative conservation probability of a motif, \( P \), the probability of each residue of a motif having its given RLC or higher can be calculated as the product of the \( P_{\text{RLC}} \) for each residue within the motif. A significance value, \( P_{\text{cons}} \), representing the probability of a given motif having that \( P_{\text{value}} \) or lower by chance, can then be calculated for the motifs \( P_{\text{value}} \) using the CDF of the uniform product distribution [Equation (1)]. Thus, the \( P_{\text{cons}} \) statistic provides a useful measure of how likely it is that this motif will have the observed degree of local conservation (or higher) by chance. Note that it does not provide any indication of the probability of the motif itself, which is best inferred from the enrichment values.

\[
P_{\text{cons}} = \frac{(-1)^n \Gamma(n, \ln(P))}{n!} \frac{n!}{\ln(P)^n} 
\]
\( P(\text{cons}) \) is the probability of a given motif having that
\( P \)-value or higher by chance, calculated as the CDF of
the uniform product distribution, i.e. the distribution of
the product of \( n \) uniform distributions, where \( n \) is the
number of non-wildcard positions in the motif, \( P \) is the
relative conservation probability of a motif and \( \Gamma \) is the
incomplete gamma function.

Enrichment scores

Enrichment scores for motif counts are calculated for the
input motif against the reverse of the motif and a
randomly shuffled variant of the motif. The score is a
simple quotient, where the input motif count is the
divisor and the shuffled or reverse motif count is the
dividend. Enrichment scores for each GO term and
BioGRID interaction hub protein are calculated versus
the expectation provided by the whole proteome, i.e. the
number of motif occurrences in proteins with that GO
term/interaction partner divided by the expected number
of proteins, which is the total number of proteins with the
motif multiplied by the proportion of the proteome with
that GO term/interaction partner. Enrichment significance
is calculated using the Fisher’s exact test. Counts are
normalized for independence by clustering highly similar
proteins based on UniRef50 groups.

THE SLIMSEARCH 2.0 WEBSERVER

The SLiMSearch 2.0 server is available at http://bioware.
.ucd.ie/slimsearch2.html. The website is free and open to
all and there is no login requirement. The purpose of the
web server is to allow researchers to identify novel occur-
cences of user-defined SLiMs in a set of sequences. A rapid
pattern matching search is first performed to identify all
occurrences of the motif in the proteome (or a defined
subset). Pre-formatted databases are then used to rapidly
extract scores and sequence features for each occurrence
before enrichment scores are calculated. Interactive output
and visualizations permit easy exploration of returned oc-
currences of the motif and their sequence context. These
features of the web server are described in more detail in
the following sections.

Input

A motif to be compared against the search is the sole com-
pulsory input. The motif should be expressed as a regular
expression using single letter amino acid codes (e.g. R.LF
or RxLF but not Arg–x–Leu–Phe). The format allows for
ambiguity (i.e. positions that can be any residue from a set
of residues, e.g. [ILV] meaning any aliphatic residue),
flexibility (e.g. \( .\{1,3\}\), meaning a wildcard position be-
tween 1 and 3 residues in length), termini definition
(where \( \ast \) is the N-terminus and \( \ast \) is the C-terminus) and
conditional motifs (e.g. (motif1)/(motif2) meaning motif1
or motif2). Two optional filtering options are also avail-
able, restricting the protein search space to a subset of a
proteome: by GO term (in the format GO:0005868) to
restrict the search to a particular ontology and similarly,
to a set of proteins by UniProt accessions. For clarity,
example inputs are available above each entry box on
the input page of the web server.

Submitting jobs

Once input has been determined, clicking ‘Submit job’ will
enter the run queue. Run times will vary according to
input data size, motif complexity and the current load of
the server but are generally in the order of a few seconds.
Users can either wait for their jobs to run or bookmark
the page and return to it later, although jobs are deleted
after 21 days. The web server can also be run directly using
a URL containing the motif to be searched and (option-
ally) a list of UniProt IDs.

Output

The main output is a table of motif instances annotated
with attributes including: (i) conservation and disorder
statistics; (ii) overlapping feature, such as Pfam domains,
PDB structures, SNPs and modifications; and (iii) over-
lapping experimentally validated motifs (Figure 1). In
addition, alignments of 100 amino acid regions over-
lapping each motif occurrence can be visualized. Discovered motifs are not filtered, therefore all instances
are returned. By default, motifs are ranked based on
\( P(\text{cons}) \). Several additional tables are also returned: GO
terms which are enriched for the motif; hub proteins where
the interactors are enriched for the motif and motif count
statistics. All results are returned as tab-delimited files
and in a more visually appealing html format. Initially,
an overview of the most interesting instances and enrich-
ments are returned. More detailed data are available and
can be sorted by each attribute. Instance data can be also
filtered based on IUPred mean disorder score, IUP.

Users need to consider two separate lines of evidence
when assessing the significance or otherwise of the findings
presented. First, the motif enrichment over the reversed
and shuffled sequences gives an indication to what extent
the motifs that are provided occur by chance. If a motif
occurs 40 times and the reverse occurs 20 times, this means
that we expect that about half of the observed instances
are false positives (assuming no negative selection on
randomly occurring motifs). The user can then scroll
down the list of occurrences, and investigate the conser-
vation values, to form a judgement regarding which motifs
are most likely to be true positives. Assuming a typical
mammalian motif, it would be expected in this case that
the 20 least conserved motifs are most likely to be false
positives and the 20 most conserved are most likely to be
true positives. In many cases, the enrichment may be rela-
tively modest; \( P(\text{cons}) \) only provides guidance, rather than
proof, regarding the likelihood that a given motif occur-
rence is a true positive.

Example analysis

The web server incorporates a full example for searching
the human proteome with the manually curated, experi-
mentally validated, Dynel Light Chain binding motif
([KR],|QT: ELM entry LIG Dynel DLC8_1 (2)). A
full walkthrough for this data set is provided in the help
pages and fully interactive example output is also
Motif Hits

Top 10 by conservation (See all)

<table>
<thead>
<tr>
<th>Accession</th>
<th>Gene</th>
<th>GO ID</th>
<th>GO term</th>
<th>Proteins</th>
<th>p(con Ret)</th>
<th>UCS</th>
<th>IUP</th>
<th>Motif</th>
<th>RE Start</th>
<th>Pfam</th>
<th>POB</th>
<th>Mod</th>
<th>SHP</th>
<th>ELM</th>
<th>Other</th>
<th>Alignment</th>
</tr>
</thead>
<tbody>
<tr>
<td>G6PJD2</td>
<td>C14orf43</td>
<td></td>
<td></td>
<td></td>
<td>0.003</td>
<td>0.8</td>
<td>0.752</td>
<td>HATOT</td>
<td>P[IR] TO7</td>
<td>969</td>
<td>view</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>G9YS3P</td>
<td>RAR2</td>
<td></td>
<td></td>
<td></td>
<td>0.004</td>
<td>0.89</td>
<td>0.793</td>
<td>HGTOT</td>
<td>P[IR] TO7</td>
<td>777</td>
<td>view</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>G9YS57</td>
<td>RASGRF1</td>
<td></td>
<td></td>
<td></td>
<td>0.005</td>
<td>0.87</td>
<td>0.473</td>
<td>HATOT</td>
<td>P[IR] TO7</td>
<td>671</td>
<td>view</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>P42373</td>
<td>ARH5AP2</td>
<td></td>
<td></td>
<td></td>
<td>0.007</td>
<td>0.94</td>
<td>0.57</td>
<td>HRTOT</td>
<td>P[IR] TO7</td>
<td>438</td>
<td>view</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Q134409</td>
<td>CYTOSOL</td>
<td></td>
<td></td>
<td></td>
<td>0.01</td>
<td>0.95</td>
<td>0.583</td>
<td>HRTOT</td>
<td>P[IR] TO7</td>
<td>158</td>
<td>view</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Q45315</td>
<td>CYTOSOL</td>
<td></td>
<td></td>
<td></td>
<td>0.01</td>
<td>0.96</td>
<td>0.643</td>
<td>HRTOT</td>
<td>P[IR] TO7</td>
<td>491</td>
<td>view</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Q14571</td>
<td>CYTOSOL</td>
<td></td>
<td></td>
<td></td>
<td>0.01</td>
<td>0.89</td>
<td>0.565</td>
<td>HERTOT</td>
<td>P[IR] TO7</td>
<td>168</td>
<td>view</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Q9Y928</td>
<td>TRAF3P3</td>
<td></td>
<td></td>
<td></td>
<td>0.026</td>
<td>0.83</td>
<td>0.222</td>
<td>RFGOT</td>
<td>P[IR] TO7</td>
<td>164</td>
<td>view</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>QPN1761</td>
<td>ATP</td>
<td></td>
<td></td>
<td></td>
<td>0.034</td>
<td>0.75</td>
<td>0.34</td>
<td>IRTOT</td>
<td>P[IR] TO7</td>
<td>908</td>
<td>view</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>P46013</td>
<td>ANTX</td>
<td></td>
<td></td>
<td></td>
<td>0.043</td>
<td>0.54</td>
<td>0.652</td>
<td>HLRTOT</td>
<td>P[IR] TO7</td>
<td>2017</td>
<td>view</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
</tbody>
</table>

Top 5 enriched GO terms by enrichment significance (See all)

<table>
<thead>
<tr>
<th>p</th>
<th>GO id</th>
<th>GO term</th>
<th>Proteins</th>
</tr>
</thead>
<tbody>
<tr>
<td>1.5e-06</td>
<td>113</td>
<td>GO:0000218</td>
<td>3 proteins</td>
</tr>
<tr>
<td>5.8e-05</td>
<td>5.37</td>
<td>GO:0007165</td>
<td>13 proteins</td>
</tr>
<tr>
<td>0.000179</td>
<td>26.5</td>
<td>GO:0012298</td>
<td>3 proteins</td>
</tr>
<tr>
<td>0.000238</td>
<td>8.83</td>
<td>GO:0008584</td>
<td>5 proteins</td>
</tr>
<tr>
<td>0.000895</td>
<td>42.9</td>
<td>GO:0007120</td>
<td>2 proteins</td>
</tr>
</tbody>
</table>

Top 5 enriched interactors by enrichment significance (See all)

<table>
<thead>
<tr>
<th>p</th>
<th>Interactors</th>
<th>Protein</th>
</tr>
</thead>
<tbody>
<tr>
<td>0.0000</td>
<td>5</td>
<td>Dynamin light chain 1</td>
</tr>
<tr>
<td>0.0000</td>
<td>3</td>
<td>Dynamin light chain 2, cytoplasmic</td>
</tr>
<tr>
<td>0.0001</td>
<td>3</td>
<td>D8 ubiquitin-protein ligase Rbf2</td>
</tr>
<tr>
<td>0.0001</td>
<td>2</td>
<td>Hsp70-binding protein 1</td>
</tr>
<tr>
<td>0.0001</td>
<td>2</td>
<td>Dynamin light chain Tctex-1type 3</td>
</tr>
</tbody>
</table>

Motif Statistics

<table>
<thead>
<tr>
<th>Type</th>
<th>Motif</th>
<th>Enrichment Instances</th>
<th>Proteins</th>
<th>Dataset</th>
<th>Size</th>
</tr>
</thead>
<tbody>
<tr>
<td>Motif</td>
<td>PIR</td>
<td>7</td>
<td>161</td>
<td>153</td>
<td>20266</td>
</tr>
<tr>
<td>Reverse TQT</td>
<td>PIR</td>
<td>7</td>
<td>154</td>
<td>152</td>
<td>20266</td>
</tr>
<tr>
<td>Shuffle</td>
<td>PIR</td>
<td>7</td>
<td>158</td>
<td>150</td>
<td>20266</td>
</tr>
</tbody>
</table>

Figure 1. Main results page. In addition to individual statistics for the top ten motif occurrences, the default results page displays motif counts, GO enrichment and protein interactor enrichment for all occurrences. If no enriched GO terms and/or interactors have been found, these sections will be blank.

Getting help

SLiMSearch 2.0 is supported by an extensive help section, including a quickstart guide and walkthrough with screenshots. Example input files are provided and example input data can be loaded into the input forms. Fully interactive example output (corresponding to running the example input with default parameters) is clearly linked from the help pages (See ‘Example analysis’ section).

FUTURE WORK

Currently, only human proteome searches are available but other proteomes will be added with time. A selection of model organisms will be added in the near future.

CONCLUSION

There are many sources of de novo motifs, including experimental approaches such as mutagenesis and peptide arrays or phage display. With recent developments in experimental technologies for determining protein–protein interaction networks and computational techniques for predicting interaction motifs from them, the number of putative SLiMs is likely to increase dramatically in the next few years. SLiMSearch 2.0 represents a valuable tool for the annotation of such motifs. In addition to de novo motifs, the server is useful for finding candidate occurrences of established SLiMs, including those found in motif databases such as ELM (2) and MiniMotif Miner (3). Often, the definition of these motifs is not conclusive and so there are also times when it is useful to search using a specific variant or a relaxed motif definition. For many known SLiMs, we currently only have annotated occurrences for a restricted set of taxonomic groups (2) but, due to their short and degenerate nature, they often evolve convergently (31). As the number of full proteomes continues to increase, the SLiMSearch 2.0 server will enable the identification of SLiMs in new taxa, helping to shed light on the breadth and depth of functional SLiMs. The SLiMSearch 2.0 server is available at: http://bioware.ucd.ie/slimsearch2.html.

FUNDING

Science Foundation Ireland (08/IN.1/B1864) and the University of Southampton; European Molecular Biology Laboratory [EMBL Interdisciplinary Postdoc (EIPOD) fellowship to N.E.D.] Funding for open access charge: The University of Southampton.

Conflict of interest statement. None declared.

REFERENCES

3. Bailey, T.L., Boden, M., Buske, F.A., Frith, M., Grant, C.E.,
5. de Castro, E., Sigrist, C.J., Gattiker, A., Bulliard, V.,
6. Gutman, R., Berezin, C., Wollman, R., Rosenberg, Y. and
7. Bateman, A., Coin, L., Durbin, R., Finn, R.D., Hollich, V.,
8. Sigrist, C.J., Cerutti, L., de Castro, E., Langendijk-Genevaux, P.S.,
11. Via, A., Gould, C.M., Gemund, C., Gibson, T.J. and Helmer-
13. Bailey, T.L., Boden, M., Buske, F.A., Frith, M., Grant, C.E.,
14. Via, A., Gould, C.M., Gemund, C., Gibson, T.J. and Helmer-
21. Via, A., Gould, C.M., Gemund, C., Gibson, T.J. and Helmer-
23. Ashburner, M., Ball, C.A., Blake, J.A., Botstein, D., Butler, H.,